GENERATION AS AUXILIARY SUPERVISION: ENHANCING VISUAL UNDERSTANDING AT ZERO INFERENCE OVERHEAD VIA DECOUPLED EMBEDDING PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

While Multimodal Large Language Models (MLLMs) have achieved remarkable progress, visual understanding and generation are typically treated as divergent objectives. Existing unified frameworks often rely on discrete visual tokenization or diffusion objectives whose generative targets differ from the continuous representations consumed by visual understanding models, making direct transfer to enhance existing pretrained MLLMs non-trivial. In this work, we present **GAS**, a generation-guided training framework that reinterprets visual generation as auxiliary supervision for representation learning. Concretely, GAS adapts **Next Embedding Prediction (NEP)** as a cross-modal generation paradigm within a decoupled Mixture-of-Transformers (MoT) architecture. By maintaining a shared lower trunk and parallel upper layers, GAS lets generation losses enrich the shared visual pathway with finer spatial precision and stronger visual retention while shielding the upper understanding layers from direct generation gradients. To maximize this synergy, we further construct highly correlated generation tasks that demand deep cognitive grounding rather than generic synthesis alone. Across model scales and training stages, GAS improves aggregate multimodal understanding, with its most reliable gains on perception and spatial comprehension. Crucially, because the auxiliary generation branch is discarded after training, these gains incur **zero inference overhead**. Extensive controlled comparisons and representation-level analyses further clarify when and why generation-guided training benefits understanding, and demonstrate the feasibility of generation-guided training as a practical route to stronger multimodal understanding.

1 INTRODUCTION

Multimodal large language models (MLLMs) have achieved remarkable progress in visual understanding tasks, including image captioning, visual question answering, and visual grounding Li et al. (2024a); Bai et al. (2025a); Zhu et al. (2025). The prevailing paradigm encodes visual inputs through a vision encoder and projects the resulting representations into token space of large language model (LLM), which then performs reasoning and generates textual responses in an autoregressive manner Liu et al. (2023); Radford et al. (2021). Although this architecture has proven remarkably effective for semantic-level comprehension, its training objective – next-token prediction (NTP) over text – provides only indirect supervision for visual perception. The model is trained to talk about images, yet never to learn from them because its text objective does not directly supervise pixel- or region-level visual structure.

This text-only supervision introduces limitations from two complementary perspectives. On one hand, natural language inherently suffers from an **expressiveness bottleneck**: it cannot precisely encode fine-grained visual details such as exact spatial relationships, pixel-level object boundaries, physical interaction dynamics, or subtle appearance variations Shao et al. (2024); Shiri et al. (2024). When the sole training signal is text prediction, the model's visual representations are shaped only by what language can articulate, leaving the richer geometric, structural, and relational information latent and underutilized. On the other hand, there exist abundant **visually intensive tasks** like segmentation and image editing. Existing VLM architectures are structurally unable to consume such data for training, thereby forfeiting a vast reservoir of fine-grained visual supervision that could otherwise cultivate

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

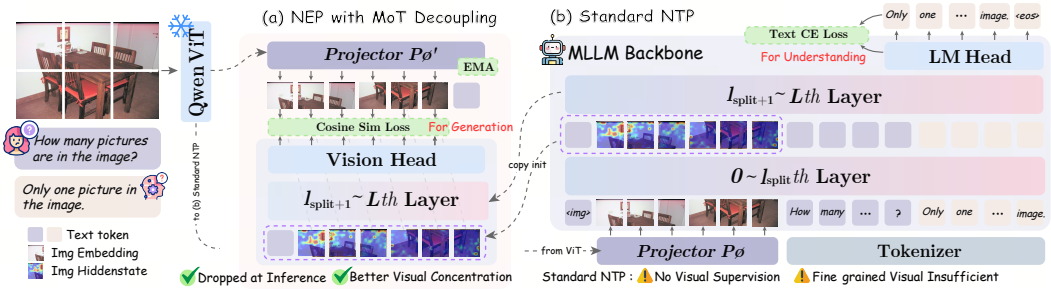


Figure 1: **Overview of our generation-guided training framework.** Given multimodal inputs, the understanding branch (b) processes visual tokens from shared vision encoder and predicts text via standard NTP paradigm. The generation branch (a), operating on independent MoT layers, receives intermediate hidden states from understanding backbone and predicts target image embeddings via Next Embedding Prediction (NEP). At inference time, the generation branch is entirely discarded.

deeper perception and complex visual understanding capabilities. Furthermore, empirical evidence reveals that visual information progressively attenuates in deeper LLM layers during reasoning Zhang et al. (2024); Ganz et al. (2024), precisely because the text-centric objective provides no mechanism to encourage persistent retention of visual features throughout the inference chain.

A natural hypothesis emerges: if a model is additionally trained to predict visual content, such as pixel-level segmentation masks, edited images, or visual reasoning traces, then its internal representations may be encouraged to retain richer and more precise visual information that can benefit understanding. This intuition has motivated the recent wave of unified multimodal models (UMMs) that jointly train understanding and generation Team (2024); Xie et al. (2025); Zhou et al. (2025). However, directly adopting the UMM paradigm to enhance understanding exposes two unresolved tensions:

First, current UMMs retain generation-side parameters at inference time Deng et al. (2025); Shen et al. (2025), incurring latency and memory costs that hinder deployment. Moreover, generation objectives in these systems are typically optimized for synthesis quality rather than for maximizing benefit to comprehension Fu et al. (2026b). Understanding performance can therefore plateau or even degrade relative to MLLMs of comparable scale, suggesting that simple joint training does not reliably transfer generative learning into understanding gains.

Second, existing work typically reports aggregate benchmark numbers after introducing generation data, without dissecting the underlying mechanisms. Critical questions remain open: Do generation and understanding objectives inherently conflict? Which intermediate representations are most amenable to generative supervision for "feeding back" into understanding performance? Which types of generation tasks enhance which facets of comprehension? And through what pathways does generation training confer its advantages?

In this work, we propose a generation-guided training framework that resolves both tensions. Our core idea is to leverage generation tasks purely as training-time auxiliary supervision for understanding, rather than as an end goal. Architecturally, as shown in Figure 1, we adopt a Mixture-of-Transformers (MoT) design that introduces a dedicated generation branch with independent transformer parameters alongside the understanding branch, so that generation gradients refine shared visual representations without directly perturbing understanding-side parameters. The generation branch employs the Next Embedding Prediction (NEP) paradigm to predict continuous visual embeddings autoregressively in the representation space consumed by the LLM backbone, without requiring a pixel decoder or discrete visual de-tokenizer. At inference time, the generation branch is entirely discarded, yielding strictly zero additional cost during inference. This deployment benefit is distinct from training cost: GAS uses approximately 11.6% more GPU-hours than the corresponding baseline.

On the data side, we systematically construct about 10M generation samples across five primary categories and 15 subtasks, covering a spectrum from pixel-level perception to high-level reasoning, and develop an automated synthesis pipeline that produces high-quality training data without manual annotation. This framework allows us to conduct systematic analyses to elucidate how generation enhances complex visual understanding which leads to several intriguing findings on generation-understanding conflict, task-specific ablation and representation-level diagnostics.

We finally present the **GAS** model family, validating the framework on 2B- and 4B-parameter backbones. GAS improves aggregate visual understanding at both scales, while the per-benchmark transfer is task-dependent. Our main contributions are as follows:

- We combine continuous NEP supervision with a removable MoT generation branch, transferring generation-side learning to the deployed understanding model at zero inference overhead.
- We construct diverse generation task categories with an automated data synthesis pipeline, and show that tasks with stronger latent correlations to understanding deliver the largest gains, while combining them yields further, complementary improvements on perception- and reasoning-oriented benchmarks.
- We provide systematic analyses of parameter isolation, layer-wise supervision injection, per-task contribution, representation-level diagnostics, matched-budget controls and repeated-run statistics, clarifying when and how generation enhances understanding.

2 METHODOLOGY

Our framework is designed around a single guiding principle: generation tasks should serve as training-time auxiliary supervision that enriches the model's visual representations, without altering the inference-time architecture or cost. To realize this, we introduce three tightly integrated components: (1) Next Embedding Prediction (NEP) paradigm: formulates image generation as autoregressive embedding prediction within the same representation space used for understanding; (2) Mixture-of-Transformers (MoT) architecture: decouples generation and understanding parameters while preserving a shared visual representation pathway; and (3) systematic multi-type generation data construction strategy that maximizes the complementary benefits of diverse visual tasks.

2.1 NEXT EMBEDDING PREDICTION (NEP)

Existing unified models typically adopt either discrete visual tokenization Wang et al. (2026c); Wu et al. (2025) or diffusion-based generation Yang et al. (2026); Xie et al. (2025). Both approaches introduce representation spaces fundamentally disjoint from the continuous embeddings consumed by the understanding branch, severely limiting how much generation training can benefit the shared visual representations Team et al. (2026).

To bridge this gap, we propose **Next Embedding Prediction (NEP)** as a cross-modal generation paradigm. Recent literature Xu et al. (2025) has demonstrated that predicting continuous embeddings autoregressively serves as an exceptionally strong visual learner. Building upon this insight, our core motivation is to formulate image generation as autoregressive embedding prediction within the exact same continuous representation space used by the LLM backbone.

This formulation anchors the generative target directly in the continuous input space of the language model, and preserves a unified representation manifold for both perception and generation. Unlike discrete tokenization which maps images into a separate categorical codebook, or diffusion objectives defined in a specialized synthesis-oriented latent space, NEP avoids a separate target-to-LLM translation interface. The design is therefore intended to make auxiliary visual supervision more directly compatible with the representations used for downstream understanding; Section 4 tests this claim against fixed-architecture and related-objective controls.

Contemporaneous UniHetero Chen et al. (2025a) likewise finds that autoregressing on LLM input embeddings is effective at large data scale. NEP is differentiated by using this representation as the target for instruction-conditioned cross-image prediction across diverse generation tasks, and by coupling it to an asymmetric MoT branch whose upper-layer gradients are isolated from the understanding path and whose parameters are removed after training.

Formulation and Supervision Target. Unlike single-image self-supervised pretraining Xu et al. (2025), NEP operates in a conditional, multi-task setting. Given a multimodal context containing an input image I and diverse linguistic instructions (e.g., for segmentation or grounding), let $\mathbf{x}_{\text{ctx}}$ denote the concatenated sequence of text tokens and source image embeddings. For a generative task with a target image I^{tgt}, the model autoregressively predicts the target embedding sequence $\hat{\mathbf{z}}^{\text{tgt}}$.

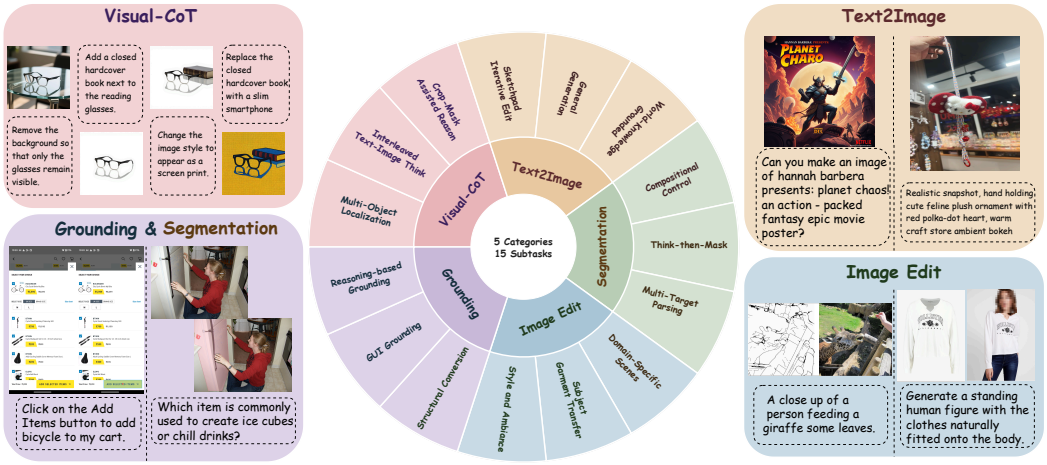


Figure 2: **Overview of generation training data.** Five primary task categories spanning 15 subtasks, jointly covering the fine-grained visual capabilities targeted by our generation-guided training.

Crucially, rather than introducing an external reconstruction decoder to predict raw pixels, we extract ground-truth embeddings using the identical vision encoder followed by a visual-language projector. To ensure the target resides in the exact LLM input space, the target embeddings are defined as $\mathbf{z}^{\text{tgt}} = \text{Projector}(\text{ViT}(I^{\text{tgt}}))$. The model predicts each embedding conditioned on the context and previously generated embeddings:

$$\hat{z}_i^{\text{tgt}} = f_{\text{gen}}\big(\mathbf{x}_{\text{ctx}}, \hat{z}_{<i}^{\text{tgt}}; \Theta_{\text{gen}}\big), \tag{1}$$

where f_{gen} denotes the generative forward pass of the network parameterized by Θ_{gen}. To optimize this objective, both the predicted and target embeddings are first ℓ_2-normalized. The generation loss is then computed as the cosine distance in this continuous representation space:

$$\mathcal{L}_{\text{gen}} = \frac{1}{N}\sum_{i=1}^{N}\left(1 - \frac{\hat{z}_i^{\text{tgt}} \cdot z_i^{\text{tgt}}}{\|\hat{z}_i^{\text{tgt}}\|_2 \|z_i^{\text{tgt}}\|_2}\right). \tag{2}$$

By optimizing directly within the understanding interface rather than pursuing photorealistic pixel synthesis, we explicitly prioritize the enrichment of deep structural representations over visual fidelity.

Target Stabilization via EMA. A critical challenge in predicting representations projected into the LLM space is that the active visual projector parameters are continuously updated during joint training Chen et al. (2025a). Utilizing a dynamic projector to extract targets can lead to supervision drift and representation collapse. To prevent this, we maintain a *target projector* whose weights are updated via an Exponential Moving Average (EMA) of the active projector's weights. The stable target embeddings $\mathbf{z}^{\text{tgt}}$ are thus computed by passing the target image through the frozen ViT followed by this EMA target projector. Correspondingly, a dedicated vision-head is applied to the output hidden states of the LLM's generation branch to predict these stable targets. As detailed in Section 2.2, the decoupled architecture with this robust NEP formulation safely injects precise spatial information into the shared representations, and further mitigates direct interference with the upper understanding layers.

2.2 MoT Decoupled Architecture

A straightforward approach to introducing generation supervision is to add the NEP loss directly onto the shared LLM backbone. However, our experiments (Section 4) and prior works reveal that this strategy degrades understanding capability Deng et al. (2025); Li et al. (2025); Hao et al. (2025): the generative and understanding objectives impose conflicting optimization pressures, ultimately diluting the abstract semantic representations necessary for complex reasoning. To resolve this tension, we adopt the Mixture-of-Transformers (MoT) architecture Liang et al. (2025), which decouples the upper task-specific parameters while preserving a shared lower trunk for indirect knowledge transfer.

Decoupled Architecture. To preserve visual structures that are typically abstracted in deeper layers focused on semantics, the model topology bifurcates at an intermediate layer l_{split}. The understanding

branch continues to process the sequence through the remaining upper layers ($l > l_{\text{split}}$), optimized solely by the text cross-entropy loss ($\mathcal{L}_{\text{und}}$). Concurrently, we extend the backbone with a parallel generation branch comprising N_{gen} transformer layers. To guarantee a reasonable initialization and stabilize early training, the parameters of these generation layers (Θ_{gen}) are directly copied from the corresponding pre-trained layers of the understanding backbone. This branch extracts hidden states from layer l_{split} to autoregressively produce NEP predictions:

$$\mathbf{H}_{\text{gen}} = \text{Transformer}_{\text{gen}}(\mathbf{H}^{(l_{\text{split}})}; \Theta_{\text{gen}}) \tag{3}$$

$$\hat{z}_i^{\text{tgt}} = \text{VisionHead}(\mathbf{h}_{\text{gen},i}; \Theta_{\text{head}}), \tag{4}$$

where $\mathbf{H}^{(l_{\text{split}})}$ denotes the sequence of hidden states at layer l_{split}, and $\mathbf{h}_{\text{gen},i}$ is the i-th token representation in the generation output $\mathbf{H}_{\text{gen}}$. The generation branch is parameterized by Θ_{gen} and Θ_{head}, supervised entirely by the NEP loss ($\mathcal{L}_{\text{gen}}$).

Mechanism of Synergy. A critical question arises: how can generative training enhance understanding if their parameters diverge? The key lies in the asymmetric gradient flow across the shared first l_{split} layers. This partial decoupling induces two powerful, complementary effects:

(1) Direct Refinement of Shared Representations. During joint training, gradients from $\mathcal{L}_{\text{gen}}$ backpropagate through the generation branch into the visual projection and the shared trunk ($l \leq l_{\text{split}}$). This forces these lower layers to produce intermediate states $\mathbf{H}^{(l_{\text{split}})}$ that are enriched with precise, fine-grained visual details. Because $\mathbf{H}^{(l_{\text{split}})}$ acts as the direct input to the upper understanding branch, this visual enrichment is seamlessly inherited.

(2) Implicit Adaptation of Upper Layers. While the upper understanding layers ($l > l_{\text{split}}$) remain shielded from generative gradients, they are continuously fed the visually enriched representations from the shared trunk. Consequently, they are compelled to adapt under $\mathcal{L}_{\text{und}}$ to fully exploit these high-quality features. Empirically (Section 4), this adaptation manifests as sharpened attention heads that reliably anchor onto question-relevant visual regions, thereby enhancing evidence retrieval during multi-step reasoning.

Training and Inference. The overall objective minimizes the weighted sum $\mathcal{L} = \mathcal{L}_{\text{und}} + \lambda \cdot \mathcal{L}_{\text{gen}}$. Crucially, at inference time, the entire generation branch (Θ_{gen} and Θ_{head}) is discarded. Only the understanding branch performs the forward pass, rendering the inference cost, latency, and memory footprint strictly identical to the standard baseline model.

2.3 Generation Task Construction

Prior unified multimodal works typically introduce generation data in the form of generic text-to-image synthesis or simple editing, treating generation merely as a self-contained skill. Under such configurations, understanding performance gains are often negligible or even negative. We argue that generic synthesis tasks, such as "generate an image of a cat", provide little training signal relevant to complex visual reasoning, as they merely demand a superficial alignment between basic nouns and visual concepts.

The core insight of our framework is that **generation can better enhance understanding when the generative act inherently depends on deep multimodal comprehension.** Rather than treating generation as a straightforward rendering process, we purposefully formulate it as the terminal output of a complex cognitive task. If a model must execute precise spatial localization, compositional scene analysis, or multi-step logical deduction just to figure out what or where to generate, the generation objective acts as a powerful forcing function. It compulsorily drives the network to develop the exact fine-grained perceptual and reasoning capabilities that pure text supervision fails to provide.

Guided by this principle, we select candidate generation tasks according to their latent cognitive dependency on the target understanding capabilities. We construct five generation task categories with 15 subtasks in total, including Grounding, Segmentation, Image Editing, Visual Chain-of-Thought, and Text-to-Image (illustrated in Figure 2). These tasks are designed to cultivate fine-grained spatial reasoning, instance-level discrimination, compositional scene understanding, multi-step visual reasoning, and world knowledge injection, respectively. Detailed descriptions of all task categories and subtasks are provided in Appendix B.

Crucially, our ablation studies (Section 4) confirm that these task categories provide complementary supervision signals. Empirically, on Counting & Spatial benchmarks the gain from combining all

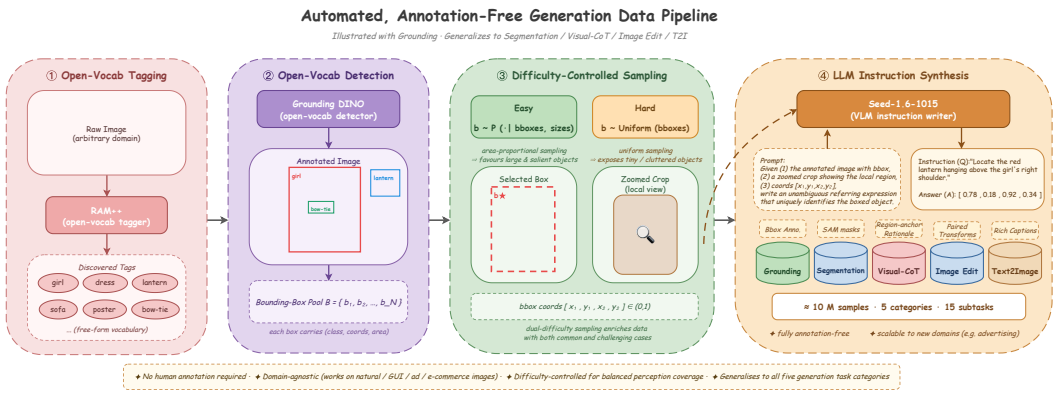


Figure 3: **Generation task construction pipeline.** Illustration of our automated, annotation-free data construction pipeline, taking grounding as a representative example.

categories tends to exceed the best single-task gain, suggesting that diverse, comprehension-correlated generation tasks are not redundant; rather, they jointly cover different facets of fine-grained visual structure that no single task fully captures.

Building on this finding, we develop a scalable data synthesis framework to produce high-quality generation data across diverse domains without manual annotation (illustrated in Figure 3). Taking grounding as a representative example, our automated pipeline operates in four stages: (1) open-vocabulary tagging via RAM++ to discover all salient objects; (2) open-vocabulary detection via Grounding DINO to localize each tag; (3) difficulty-controlled sampling (area-proportional for easy, uniform for hard) to select target boxes; and (4) LLM-based instruction synthesis using the annotated image and a zoomed crop to generate unambiguous referring expressions. This pipeline generalizes to arbitrary image domains and enables scaling of generation data to new scenarios (e.g., advertising) with minimal engineering effort.

Combining the above task selection principle with this automated construction and filtering pipeline, we ultimately scaled the total generation training corpus to approximately 10M samples spanning all five task categories, and mixed them with understanding data for training.

3 Experiments

3.1 Experiments Setup

Model Configuration. To ensure a controlled environment devoid of data contamination from existing VLM weights, we build our main models from scratch. Adopting the Qwen3-VL architecture Bai et al. (2025a) for the understanding branch, we utilize Qwen3-VL-ViT as a visual encoder initialized with pretrained weights, which remains frozen throughout all training stages. For the LLM backbone, we employ the 2B and 4B Qwen3 Yang et al. (2025) models. Note that we use the raw LLM weights without any prior multimodal alignment. Additionally, the ViT-LLM projector is randomly initialized, ensuring all visual-language capabilities are acquired solely through our training pipeline. The MoT generation branch consists of L_{mot} transformer layers initialized from the corresponding pre-trained layers of the understanding backbone, with hidden states extracted from the intermediate layer $l_s \approx L/2 = 14$ of the understanding backbone (where L is the total layer count).

Implementation Details. Training proceeds in two stages. In **Stage 1 (Align)**, we freeze all understanding-side parameters and train only generation branch using T2I generation data so as to activate the generation pathway and align it with the existing visual representation space without perturbing the pretrained understanding capability. In **Stage 2 (Joint)**, both understanding and generation data are used simultaneously. The understanding branch is optimized with text cross-entropy loss, while the generation branch is optimized with NEP loss under a progressive weighting schedule that linearly ramps λ from an initial value of 0.015 to the target weight 1.0 over 4k steps.

This setting allows generation gradients to refine the shared trunk while leaving the ViT frozen, thereby enabling the indirect transfer mechanism described in Section 2.2. On our HPC Cluster, the final

baseline uses 2,208 GPU-hours, whereas GAS uses 2,464 GPU-hours (+11.6%); the corresponding ablation runs use 544 and 608 GPU-hours (+11.8%). These totals include the generation branch, vision head, and EMA target projector. All three auxiliary components are removed at inference.

As for hyper-parameters, we use the AdamW Kingma & Ba (2014) optimizer with a cosine learning rate schedule, a peak learning rate of 5×10^{-5} for both understanding branch and the generation branch. We maintain a target projector EMA with a decay rate of 0.999 for training stability. Training is conducted on 32 GPUs with DeepSpeed ZeRO-2, using a global batch size of 128. The maximum sequence length is 8192 tokens.

3.2 Evaluation Setup

To comprehensively evaluate the impact of generation-guided training on complex visual understanding, we organize the evaluation into four capability dimensions: General Perception, Visual Reasoning, Counting & Spatial Understanding, and Video Understanding. Together, these dimensions assess holistic multimodal perception, multi-step visual reasoning, fine-grained spatial understanding, and temporal reasoning. The complete list of benchmarks and their corresponding evaluation focuses is provided in Appendix C.

3.3 Main Results

Table 1: **Overall results** for a range of understanding-only MLLMs, Unified Models, baselines and our GAS family. * denotes metrics taken directly from public reports rather than reproduced by us.

Model	#Params	MME test	MMMU val	BLINK test	RealWorldQA test	CharXiv DS	CharXiv RS	DynaMath test	MathVision mini	MathVista mini	LogicVista mini	VisuLogic test	CountBenchQA test	CV-Bench 2D	CV-Bench 3D	Video-MME test	MVBench test
Understanding-only MLLMs																	
Thyme-VL* Zhang et al. (2026b)	7B	-	-	56.1	70.2	-	-	-	27.6	70.0	-	23.4	-	-	-	-	-
LLaVA-v1.5* Liu et al. (2024)	7B	-	35.7	-	54.8	-	-	16.6	8.52	-	-	24.6	-	-	-	-	-
Qwen2.5-VL Bai et al. (2025b)	3B	2204	49.2	48.6	64.1	-	-	-	14.8	-	-	-	-	-	-	54.1	61.3
Qwen3-VL Bai et al. (2025a)	2B	2026	48.1	54.9	63.7	65.0	38.8	30.2	20.7	51.4	36.5	26.4	86.9	70.2	86.8	48.1	50.4
Unified Models																	
BAGEL* Deng et al. (2025)	7B+7B	2381	55.3	-	72.8	-	-	-	-	73.2	44.3	41.7	82.5	-	-	-	-
Emu3* Wang et al. (2026c)	8B	-	31.6	-	57.4	-	-	-	-	-	-	24.7	65.2	-	-	-	-
MetaMorph* Tong et al. (2025)	8B	-	41.8	-	58.3	-	-	-	-	-	-	-	-	-	-	-	48.8
Janus-Pro Chen et al. (2025c)	7B	1978*	41.0*	39.7	58.0*	49.9	24.4	24.7	15.8	45.3*	28.6	23.8	82.8	68.7	66.5	-	-
Show-o* Xie et al. (2025)	1.3B	-	26.7	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Show-o2* Xie et al. (2026)	1.5B	-	37.1	-	56.5	-	-	-	-	-	-	-	-	-	-	-	50.6
Cheers* Zhang et al. (2026a)	1.5B	-	36.0	-	60.9	-	-	-	-	50.5	-	-	-	-	-	-	-
Ours																	
GAS-Qwen3vl	2B	2003	46.6	48.0	64.8	64.4	34.3	47.9	20.4	56.4	32.7	28.6	90.1	73.2	73.8	46.2	47.8
-baseline		1986	46.9	46.7	64.1	64.0	34.6	46.2	19.1	54.4	36.0	26.8	87.7	69.9	73.2	45.2	49.0
GAS-Qwen3vl	4B	2146	57.0	56.9	70.3	78.4	40.5	55.2	27.0	68.1	37.1	23.6	90.8	75.4	83.6	54.4	52.4
-baseline		2092	54.9	57.8	70.2	77.9	40.3	57.6	25.3	65.6	37.1	26.1	90.1	75.1	82.1	55.0	54.2
GAS-Qwen3.5	4B	2384	60.8	65.8	74.9	77.0	55.1	62.9	28.0	72.9	43.0	24.8	87.9	82.2	90.4	79.9	69.2
-baseline		2381	60.6	62.9	73.6	77.0	53.9	64.0	23.4	72.6	42.7	25.0	86.4	82.1	92.1	79.4	70.3

Table 1 reports the main results of GAS, its matched understanding-only baselines, public MLLM and unified-model references across all benchmarks. The paired GAS/baseline rows are the primary evidence for attribution. All improvements of GAS are achieved with zero additional inference cost since the generation branch is discarded after training.

At the same 2B scale, GAS is highly competitive with baseline and other open-source models while showing markedly stronger reasoning and counting/spatial behavior: DynaMath increases from 46.2 to **47.9** (+1.7pp), MathVista from 54.4 to **56.4** (+2.0pp), CountBenchQA from 87.7 to **90.1**, CV-Bench-2D from 69.9 to **73.2**, and VisuLogic from 26.8 to **28.6**, all directly aligned with the spatial perception and visual reasoning capabilities targeted by our grounding, segmentation, and Visual-CoT generation tasks. Scaling to 4B, GAS establishes the strongest overall results in the unified family, leading on most of the 16 benchmarks we report (e.g., MMMU 57.0, CharXiv-DS 78.4 / RS 40.5, CountBenchQA 90.8) and surpassing significantly larger unified models such as BAGEL (7B+7B) and Emu3 (8B). The lone exception is VisuLogic at 4B (23.6), where the larger backbone shows a benchmark-specific regression that does not transfer to other reasoning suites, suggesting an idiosyncratic sensitivity of this puzzle-style benchmark to backbone scale. Together,

these results show that our generation-guided framework consistently transfers fine-grained perception and reasoning gains to the understanding branch across both scales.

Qualitative cases in Appendix F illustrate model behavior on CV-Bench, DynaMath, VisuLogic, and RealWorldQA. These selected examples show improved spatial localization, recognition of fine-grained visual details (particularly for small objects and subtle differences), and more reliable visual reasoning chains that reference image evidence. These qualitative patterns align well with the quantitative gains observed across our four evaluation dimensions, and suggest that the generation-derived supervision successfully transfers pixel-level and region-level perception capabilities to the understanding branch. In the following section, we use controlled ablations and representation-level analyses to further examine how and when generation training produces these improvements.

4 ANALYSIS

We conduct a series of controlled experiments to dissect the mechanisms underlying generation-guided training. Unless otherwise noted, analyses use the 2B-parameter model with 2.5B understanding and 2.5B generation tokens and the same evaluation suite from Section 3. We organize the investigation around five core questions: (1) Does GAS provide consistent benefits, transfer across training stages and backbone capability levels? (2) Which architecture turns generation data into transferable supervision beyond additional data, capacity, or training cost? (3) Which generation tasks contribute most to understanding gains? (4) At which layer should generation supervision be injected? (5) Which representation-level changes accompany the downstream gains?

4.1 GENERALIZATION ACROSS TRAINING STAGES

Firstly, we verify that our generation-guided framework is not limited to a specific training stage or backbone quality by applying GAS training atop understanding backbones of varying capability levels. We evaluate three settings: (i) training from scratch (the raw LLM backbone), (ii) after pretraining with large-scale multimodal data, and (iii) after directly supervised fine-tuning (SFT) based on Qwen3-VL-2B Bai et al. (2025a). Each setting compares GAS with its corresponding standard VLM baseline; matched controls for generation-data exposure and additional training budget are reported separately in Section 4.2. Results are presented in Table 2.

Table 2: **Generalization across training stages.** GAS generation-guided training yields improvements when applied to understanding backbones at different capability levels.

Stage	Configuration	Overall	Perception	Reasoning	Count&Spatial	Video	MME	MMMU	MathVista	MVBench
From Scratch	Baseline	47.25	56.55	31.38	76.92	46.60	1848	43.5	42.7	46.9
	+ GAS	**48.25**	**58.52**	**32.63**	75.15	**47.65**	**1895**	**45.8**	**47.1**	**49.5**
After Pretrain	Baseline	56.23	63.80	43.22	82.40	53.95	2169	53.2	72.3	55.1
	+ GAS	**56.71**	**63.88**	**44.12**	**82.95**	53.95	2170	51.1	71.2	54.6
After SFT	Baseline	52.65	61.32	38.23	79.82	51.40	2060	46.3	52.9	53.0
	+ GAS	**53.49**	**61.62**	**39.57**	**80.80**	**51.65**	**2096**	45.8	**53.6**	52.9

The from-scratch setting yields the largest Overall gains ($+1.00$pp overall, with $+1.97$pp on Perception and $+1.25$pp on Reasoning), but a decrease on Count&Spatial (-1.77pp), as the model benefits most from generation supervision when its visual representations are not yet well-developed. After large-scale pretraining, GAS still provides consistent improvements ($+0.48$pp overall, $+0.90$pp on Reasoning), showing that generation-guided supervision is complementary even to a strong, well-trained backbone. Most notably, applying GAS during SFT also yields a robust $+0.84$pp overall gain with simultaneous improvements on Perception ($+0.30$pp), Reasoning ($+1.34$pp), Count&Spatial ($+0.98$pp), and Video ($+0.25$pp). Across three stages spanning the full spectrum of backbone capability, GAS delivers positive transfer on virtually every dimension, confirming that the proposed generation-guided training is a stage-agnostic augmentation that benefits both weak and already-strong understanding models.

Insight 1: Generation-guided training is stage-agnostic: it provides the largest gains from scratch, complementary improvements after pretraining, and still consistent benefits after SFT, making it a plug-in augmentation across the full training pipeline rather than a stage-specific technique.

4.2 What Makes Generation Supervision Transfer?

A generation-augmented model changes several factors at once: it sees a new data stream, spends additional training compute, introduces a generation branch, and learns from a new visual objective. We disentangle these factors with the controls in Table 3. Beyond the standard 2.5B-token understanding baseline, we train a compute-matched baseline with 11% more understanding data and expose a standard VLM baseline to the same generation samples, whose image outputs contribute no loss. We then isolate the MoT architecture without generation supervision, apply NEP directly to a shared backbone, and replace NEP with position-aligned visual prediction (no shift) while keeping MoT and the training mixture fixed.

Table 3: **Disentangling data, architecture, and objective.** Matched controls separate additional training, passive exposure to generation data, MoT decoupling, and the NEP objective.

Configuration	Overall	Perception	Reasoning	Count&Spatial	Video
Understanding-only baseline	47.25	56.55	31.38	**76.92**	46.60
Baseline, +11% understanding data	47.73	57.63	34.01	71.28	45.55
Baseline, same generation data (text loss only)	47.14	56.08	33.52	71.00	46.25
MoT, no generation supervision	47.63	57.22	**34.20**	70.78	45.60
Shared backbone + NEP	46.00	54.95	32.25	70.20	45.15
MoT + no-shift visual loss	47.84	57.70	32.16	76.38	46.60
MoT + NEP (GAS)	**48.25**	**58.52**	32.63	75.15	**47.65**

Additional training alone does not explain the gain. The longer understanding-only run improves Overall only modestly, while merely mixing in the same generation samples provides no benefit because their image targets are not supervised by standard VLM training. MoT without a generation loss also improves some dimensions but does not match GAS, showing that neither data exposure nor the extra branch is sufficient by itself.

The objective and its gradient route are complementary. Applying NEP directly to the shared upper layers causes broad degradation, revealing interference between generation-specific optimization and the understanding pathway. With MoT fixed, position-aligned no-shift prediction already provides useful visual supervision, whereas autoregressive NEP yields the strongest aggregate transfer and the best Perception and Video scores. The no-shift control remains stronger on Count&Spatial, so the effect is not a uniform per-capability improvement; rather, NEP changes the transfer profile while improving the aggregate outcome. Together, these controls show that MoT protects the deployed understanding pathway and NEP supplies a task-conditioned learning signal that can enrich their shared lower representation.

Insight 2: Generation data is not supervision by itself. GAS turns it into transferable visual learning by pairing task-conditioned next-embedding prediction with an isolated upper generation pathway: NEP enriches the shared representation, while MoT prevents generation-specific optimization from rewriting the understanding model.

4.3 Per-Task Generation Contribution

We next investigate which generation tasks contribute most effectively to understanding enhancement by individually adding 200k samples from each of the five task categories to the understanding-only baseline. Table 4 presents the results, and we additionally report the combined "All" task samples configuration.

Several patterns emerge from this ablation. First, all five generation tasks improve at least one evaluation dimension, but their contribution profiles differ significantly, validating our latent task

correlation hypothesis (Section 2.3). T2I provides the highest single-task overall gain (+0.49pp) with broad improvements across perception and spatial understanding; Segmentation and Grounding strongly boost Count&Spatial (+2.00 and +2.37pp respectively, with CV-Bench 3D improving by up to 9.8pp from segmentation), consistent with the expectation that pixel-level and region-level generation tasks directly cultivate spatial perception capabilities. Visual-CoT uniquely enhances complex reasoning benchmarks: BLINK (+4.2pp) and MathVision (+1.0pp), reflecting its role in forcing the model to produce intermediate visual reasoning traces that reinforce multi-step inference. Editing yields a small Overall gain and a larger Count&Spatial improvement, but does not improve every dimension.

Table 4: **Per-task ablation.** Each row adds a single generation task category to the understanding-only baseline. The "All" row combines all five categories.

Configuration	Overall	Perception	Reasoning	Count&Spatial	Video	MME	BLINK	CV-2D/3D	MathVision	VisuLogic
Understanding only	47.63	57.22	34.20	70.78	45.60	1868	41.9	59.8/52.1	17.1	25.2
+ T2I	48.12	57.70	34.66	71.65	45.85	1820	44.5	59.3/55.3	15.8	25.1
+ Editing	47.77	56.70	34.12	72.95	45.70	1830	42.5	61.1/59.5	14.8	26.4
+ Segmentation	47.43	55.32	34.35	72.78	45.55	1750	44.1	63.6/55.9	16.4	24.2
+ Grounding	47.92	57.28	33.92	73.15	45.95	1840	44.6	61.9/58.3	13.5	23.6
+ Visual-CoT	47.86	57.80	34.26	71.45	45.20	1877	46.1	56.9/57.3	18.1	26.6
+ All (combined)	48.30	57.08	34.28	**75.72**	45.40	1851	44.9	**62.5/67.0**	**20.1**	23.6

Most notably, the "All" configuration achieves a Count&Spatial score of 75.72, higher than any single-task setting and exceeding the strongest individual contributors (Segmentation $+2.00$pp, Grounding $+2.37$pp), with a similar pattern observed on MathVision (20.1 in the combined setting versus 18.1 from the best single task). This pattern is suggestive of a complementary effect: diverse generation tasks appear to address different facets of fine-grained visual representation that no single category fully covers. The practical implication is the one we use throughout the paper: composing tasks with distinct latent correlations to understanding is a productive design lever, especially for perception- and counting-oriented capabilities.

Effect of Task Correlation on Understanding Gains. The heterogeneous gains across task categories suggest that transfer depends not only on task modality, but also on the latent correlation between generation signal and the target understanding capability, as well as on the internal characteristics of the generation task itself. To probe this factor more directly, we conduct a controlled study by fixing the T2I task and data volume and varying only prompt characteristics. Specifically, we collect 200k additional T2I samples from COYO-700M Byeon et al. (2022) dataset and compare two variants: (i) **Easy**, generic short-caption synthesis as commonly used in prior unified models, and (ii) **Rewrite**, where prompts are enriched with world knowledge, multi-object compositions, and explicit attribute/relation cues to better mirror the scene structures probed by our reasoning and perception benchmarks.

Table 5: **Effect of T2I prompt correlation with understanding tasks.** Aligning T2I prompts with the compositional, knowledge-grounded scene structures required by understanding benchmarks improves transfer at fixed task and data volume.

Metric	Baseline	Easy	Rewrite
Overall	47.63	47.87	**48.34**
Perception	57.22	56.47	**57.53**
Reasoning	34.20	34.37	**34.58**
Count&Spatial	70.78	72.47	**73.62**
Video	45.60	**46.55**	45.95

As Table 5 shows, generic T2I yields a small Overall gain ($+0.24$pp) but decreases Perception by 0.75pp. Rewritten prompts improve Overall by $+0.71$pp and all four capability aggregates relative to the same baseline, although the easy prompts retain the best Video score. This contrast shows that volume of generation supervision is not what matters, a poorly correlated generation task can act as a distractor rather than a regularizer. Once prompts are rewritten to demand the same compositional, knowledge-grounded scene structures probed by understanding benchmarks, the very same T2I category turns into a broadly beneficial signal, on par with our targeted tasks. This directly closes the loop with our latent task correlation principle (Section 2.3): even within one nominal task, transfer scales with task-level correlation, not with sample count.

Insight 3: Understanding gains scale with task correlation. Each task targets a distinct facet, like spatial tasks for counting/localization, Visual-CoT for reasoning, editing for holistic perception, and combining them tends to yield further, complementary improvements on perception- and counting-oriented benchmarks beyond any single category. Even within T2I, correlation-aware prompts unlock transfer that generic captions cannot.

4.4 Layer-wise Supervision Injection

Our MoT architecture extracts features from an intermediate layer l_s of the understanding backbone to feed into the generation branch. This design choice raises the question: at which depth should generation supervision be injected? We compare extracting hidden states from layers 8, 14 (approximately $L/2$ for our 2B, 28-layers model), and 20, as well as the effect of progressive loss warmup and unfreezing the ViT encoder. Results are shown in Table 6.

Table 6: **Layer-wise injection and training strategy ablation.** We vary the layer from which hidden states are extracted for the generation branch, and compare with progressive warmup and ViT unfreezing. Layer 14 ($\approx L/2$) with progressive warmup achieves the best overall performance.

Configuration	Overall	Perception	Reasoning	Count&Spatial	Video	MME	BLINK	CountBenchQA	CV-3D
Layer 8	47.06	56.80	31.52	75.88	45.40	1830	44.3	86.4	66.1
Layer 14	47.85	58.58	31.64	75.28	47.60	1849	47.7	86.4	64.6
Layer 14 + warmup	**48.25**	**58.52**	**32.63**	75.15	**47.65**	**1895**	45.8	**87.7**	63.6
Layer 14 + unfreeze ViT	46.20	55.80	31.31	72.60	45.25	1808	44.7	79.7	69.9
Layer 20	47.25	58.20	31.43	74.60	45.45	1855	44.0	86.0	64.2

The intermediate layer 14 (approximately $L/2$) consistently outperforms both the shallower layer 8 and the deeper layer 20 on overall metrics and perception, while layer 8 achieves the strongest Count&Spatial score (75.88), suggesting that early-layer features already encode rich spatial information that benefits from generation supervision. Layer 14 achieves the highest BLINK (47.7/45.8), indicating that deeper features carry more abstract visual-semantic information beneficial for multi-image reasoning, represents a balance point where features retain sufficient spatial detail while having undergone enough semantic abstraction to benefit broad understanding tasks.

Progressive loss warmup improves the raw layer-14 configuration by $+0.40$pp Overall and $+0.99$pp Reasoning, consistent with our motivation that the faster-converging generation loss should be introduced gradually. Conversely, unfreezing the ViT encoder reduces Overall by 1.65pp and CountBenchQA from 86.4 to 79.7 relative to the frozen layer-14 setting, indicating that the generation-side optimization pressure corrupts the pretrained visual features when allowed to propagate into the encoder, which further validating our design to keep the ViT frozen and channel generation benefits exclusively through the shared projection layer.

Insight 4: The best Overall score occurs when supervision branches near $L/2$ with progressive warmup, balancing spatial detail and semantic abstraction. Unfreezing the ViT degrades performance in this setup, suggesting that generation benefits should flow through the projection layer, not corrupt the frozen encoder.

4.5 Representation-Level Diagnostics

To understand how generation training improves understanding at the representation level, we conduct three complementary analyses: (i) visual information retention across layers, (ii) attention pattern visualization, and (iii) linear probing diagnostics.

Visual Information Retention. We measure the cosine similarity between the visual token representations at each LLM layer and the original visual features at the LLM input, tracking how much visual information is preserved as processing deepens. Figure 4 (left) shows that the baseline similarity decreases beyond layer $L/3$ whereas the GAS model retains higher similarity through the middle and deep layers, demonstrating that generation supervision encourages the network to retain visual features rather than allowing them to be "squeezed out" by language-dominant processing. The

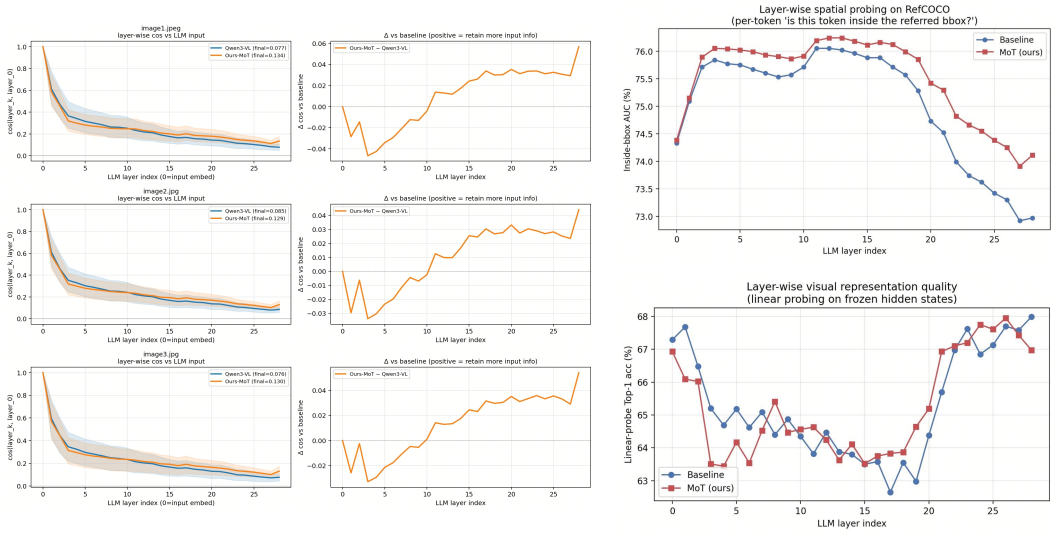


Figure 4: **Visual information retention and layer-wise linear probing.** Left: cosine similarity between per-layer visual tokens and input visual features. Right top / bottom: linear probing accuracy on RefCOCO (region-level) and ImageNet (global). The models are comparable on ImageNet, while GAS's representations become increasingly more linearly separable on RefCOCO in deeper layers.

generation branch, by requiring accurate embedding prediction from intermediate representations, creates an implicit incentive for the backbone to preserve visual information at those layers, which then propagates benefit to the understanding branch operating on the same representations.

Layer-wise Linear Probing. To quantify the discriminative quality of visual representations at each layer, we extract hidden states corresponding to visual tokens and train linear classifiers on ImageNet Deng et al. (2009) (global classification) and RefCOCO Mao et al. (2016); Kazemzadeh et al. (2014) (region-level referring expression comprehension). As shown in Figure 4 (right), the two models perform comparably on ImageNet across all layers. On RefCOCO, GAS exhibits a growing advantage that widens consistently from layer 8 onward, ultimately achieving substantially higher probing accuracy at the final layer. This pattern reveals that generation training specifically enriches fine-grained, region-level visual representations rather than generic semantic features, aligning with GAS's strongest downstream gains on spatial and grounding evaluations. The widening gap at deeper layers further corroborates the retention analysis: because GAS preserves more spatial visual information in deep layers, its representations remain useful for fine-grained tasks even after extensive language-driven processing.

Attention Visualization. We visualize attention maps from representative understanding CV-Bench example at different LLM layers to examine where the model "looks" during reasoning. More examples are shown in Appendix E. As illustrated in Figure 5, the baseline model's attention becomes more diffuse in deeper layers—by layer 20, attention is scattered broadly across the image with no clear focus on task-relevant regions. In contrast, GAS maintains sharp, focused attention on question-relevant regions throughout the network depth. For instance, on a CV-Bench counting question ("How many pictures are in the image?", shown in Boxes 1), the baseline incorrectly answers 0 with diffuse attention, while our model correctly identifies the single picture with attention concentrated on the relevant region. This attention sharpening directly explains the mechanism by which generation training improves understanding: by requiring the generation branch to predict precise visual embeddings, the shared representations develop stronger spatial selectivity that carries over to understanding-time inference.

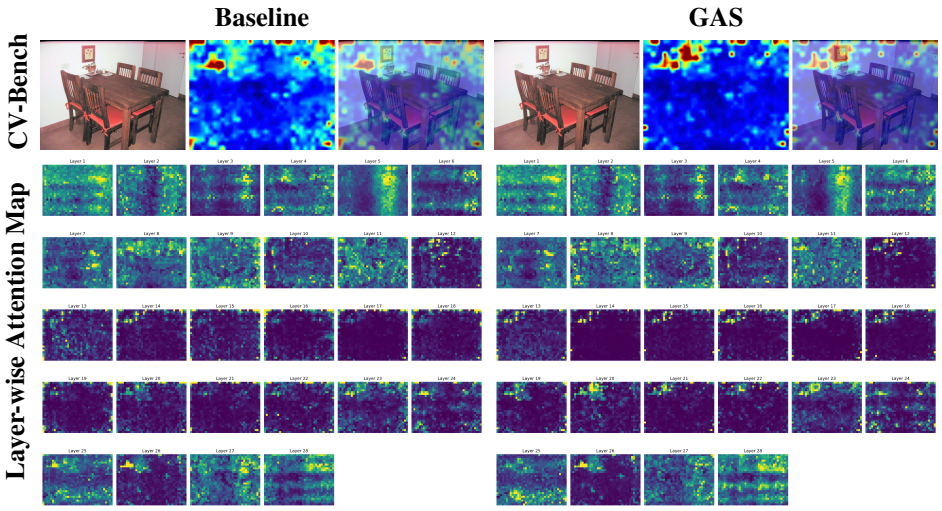


Figure 5: **Layer-wise attention visualization on representative understanding samples.** The first row compare visual-token attention for the baseline and GAS model on CV-Bench-2D Tong et al. (2024). The bottom row shows the full layer-by-layer attention evolution.

Qualitative Case 1: CV-Bench-2D (Attention Analysis)

Question: How many pictures are in the image? A. 1 B. 3 C. 2 D. 0
Ground Truth: A
Baseline: **<think>** The image shows a room with a wooden table and chairs; no separate pictures appear, so the answer is D (0). **</think>** Answer: **D ✗**
GAS: **<think>** The image is itself a single coherent scene (a dining setup); among the choices, A (1) corresponds to the single picture present. **</think>** Answer: **A ✓**

Insight 5: Generation training enhances understanding through three complementary representational mechanisms: (i) improved visual information retention in deep layers, preventing the progressive attenuation observed in text-only training; (ii) enriched fine-grained spatial representations (validated by linear probing on RefCOCO); and (iii) sharper, task-focused attention patterns that persist throughout the reasoning chain.

Text-only Reasoning capability Mildly Improved. A natural concern is whether generation supervision harms the LLM backbone's pure-text reasoning ability. We compare GAS with the understanding-only baseline on **ZebraLogic** Lin et al. (2025) (constraint-satisfaction logic puzzles) and **MMLU-Redux** Gema et al. (2025) (a re-annotated, less noisy subset of MMLU). Table 7 shows higher point estimates for GAS on both benchmarks: ZebraLogic puzzle accuracy increases by +2.8pp (with consistent improvements on small/medium puzzles and cell-level accuracy), and MMLU-Redux improves by +1.23pp. This indicates that generation-guided multimodal training does not act as a regularizer that erodes language reasoning; on the contrary, the joint optimization mildly benefits language-only capability, likely because the visual generation branch absorbs visual-specific gradient pressure that would otherwise distort the LLM's text representations.

Table 7: **Text-only reasoning evaluation.** GAS preserves and mildly improves pure-text reasoning on ZebraLogic and MMLU-Redux, indicating that generation supervision does not interfere with the LLM's language capability.

Model	ZebraLogic			MMLU-Redux
	Puzzle Acc	Cell Acc	Small Acc	Acc
Qwen3-VL baseline	18.0	12.74	52.19	52.66
+ GAS (Ours)	**20.8**	**14.79**	**56.56**	**53.89**

5 Conclusion and Future Work

We present GAS, a generation-guided training framework that improves aggregate multimodal understanding by using visual generation purely as training-time auxiliary supervision. Concretely, NEP provides a generation target aligned with the LLM input space, while the removable MoT branch mitigates direct optimization conflict and preserves zero inference overhead. Across diverse understanding benchmarks, GAS delivers consistent gains, while representation analyses associate the gains with better visual retention, more focused attention, and task-dependent transfer shaped by generation-task correlation.

These results suggest that visual generation is most effective not as an inference-time capability to preserve, but as a carefully structured supervisory signal for building stronger understanding models. We believe this work establishes generation-guided training as a principled and practical paradigm for advancing multimodal understanding. Looking forward, several promising directions emerge: scaling the framework to longer-context scenarios, extending the NEP paradigm to video-level temporal generation tasks, exploring adaptive task weighting strategies that dynamically balance generation categories during training, and investigating whether the insights on generation-understanding synergy generalize to other modalities such as audio and 3D perception.

AI use statement

(This section is **required** and does not count toward the page limit.)

In this work, we used generative AI tools for [tasks with required disclosure]. We have not used generative AI tools for [other tasks with required disclosure], and [the rest of the required disclosure tasks] are not applicable to this work. Additionally, we used generative AI tools for [tasks with recommended disclosure]. We have reviewed all AI-assisted work. [Elaborate. For example, "we checked LLM-generated research ideas for potential plagiarism through a manual literature survey", "LLM-generated code was verified and tested for correctness by 2 authors", etc.]. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

See the ICLR 2027 AI Policy for Authors for more details. This statement should not be more than 1 page.

Ethics statement

(This section is **recommended** and does not count toward the page limit.)

If authors feel that their paper submission raises questions regarding the Code of Ethics, they are encouraged to include a paragraph of Ethics Statement (at the end of the main text before references) to address potential concerns where appropriate. Topics include, but are not limited to, studies that involve human subjects, practices to data set releases, potentially harmful insights, methodologies and applications, potential conflicts of interest and sponsorship, discrimination/bias/fairness concerns, privacy and security issues, legal compliance, and research integrity issues (e.g., IRB, documentation, research ethics). This statement should not be more than 1 page.

Reproducibility statement

(This section is **recommended** and does not count toward the page limit.)

It is important that the work published in ICLR is reproducible. Authors are strongly encouraged to include a paragraph-long Reproducibility Statement at the end of the main text (before references) to discuss the efforts that have been made to ensure reproducibility. This paragraph should not itself describe details needed for reproducing the results, but rather reference the parts of the main paper, appendix, and supplemental materials that will help with reproducibility. For example, for novel models or algorithms, a link to an anonymous downloadable source code can be submitted as supplementary materials; for theoretical results, clear explanations of any assumptions and a complete proof of the claims can be included in the appendix; for any datasets used in the experiments, a complete description of the data processing steps can be provided in the supplementary materials. Each of the above are examples of things that can be referenced in the reproducibility statement.

Author Contributions

If you'd like to, you may include a section for author contributions as is done in many journals. This is optional and at the discretion of the authors.

Acknowledgments

Use unnumbered third level headings for the acknowledgments. All acknowledgments, including those to funding agencies, go at the end of the paper.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. Advances in neural information processing systems, 35:23716–23736, 2022.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL Technical Report, November 2025a. URL http://arxiv.org/abs/2511.21631. arXiv:2511.21631 [cs].

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL Technical Report, February 2025b.

Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, Thomas Unterthiner, Daniel Keysers, Skanda Koppula, Fangyu Liu, Adam Grycner, Alexey Gritsenko, Neil Houlsby, Manoj Kumar, Keran Rong, Julian Eisenschlos, Rishabh Kabra, Matthias Bauer, Matko Bošnjak, Xi Chen, Matthias Minderer, Paul Voigtlaender, Ioana Bica, Ivana Balazevic, Joan Puigcerver, Pinelopi Papalampidi, Olivier Henaff, Xi Xiong, Radu Soricut, Jeremiah Harmsen, and Xiaohua Zhai. PaliGemma: A versatile 3B VLM for transfer. arXiv preprint arXiv:2407.07726, 2024.

Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. Coyo-700m: Image-text pair dataset. https://github.com/kakaobrain/coyo-dataset, 2022.

Fengjiao Chen, Minhao Jing, Weitao Lu, Yan Feng, Xiaoyu Li, and Xuezhi Cao. Unihetero: Could generation enhance understanding for vision-language-model at large data scale? arXiv preprint arXiv:2512.23512, 2025a.

Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, et al. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. arXiv preprint arXiv:2505.09568, 2025b.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. arXiv preprint arXiv:2501.17811, 2025c.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. Advances in neural information processing systems, 36:49250–49267, 2023.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. arXiv preprint arXiv:2505.14683, 2025.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition, pp. 248–255. Ieee, 2009.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 24108–24118, 2025.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. Advances in Neural Information Processing Systems, 38, 2026a.

Fengyi Fu, Mengqi Huang, Shaojin Wu, Yunsheng Jiang, Yufei Huo, Hao Li, Yinghang Song, Fei Ding, Jianzhu Guo, Qian He, Zheren Fu, Zhendong Mao, and Yongdong Zhang. Lance: Unified multimodal modeling by multi-task synergy. arXiv preprint arXiv:2605.18678, 2026b.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. In European Conference on Computer Vision, pp. 148–166. Springer, 2024.

Roy Ganz, Yair Kittenplon, Aviad Aberdam, Elad Ben Avraham, Oren Nuriel, Shai Mazor, and Ron Litman. Question aware vision transformer for multimodal reasoning. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 13861–13871, 2024.

Aryo Pradipta Gema, Joshua Ong Jun Leang, Giwon Hong, Alessio Devoto, Alberto Carlo Maria Mancino, Rohit Saxena, Xuanli He, Yu Zhao, Xiaotang Du, Mohammad Reza Ghasemi Madani, et al. Are we done with mmlu? In Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pp. 5069–5096, 2025.

Google. Introducing Gemini 2.0: our new AI model for the agentic era, December 2024. URL https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/.

Google. Gemini 3 flash: frontier intelligence built for speed, 2025.

Jitai Hao, Hao Liu, Xinyan Xiao, Qiang Huang, and Jun Yu. Uni-x: Mitigating modality conflict with a two-end-separated architecture for unified multimodal models. arXiv preprint arXiv:2509.24365, 2025.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp. 787–798, 2014.

Fucai Ke, Joy Hsu, Zhixi Cai, Zixian Ma, Xin Zheng, Xindi Wu, Sukai Huang, Weiqing Wang, Pari Delir Haghighi, Gholamreza Haffari, et al. Explain before you answer: A survey on compositional visual reasoning. arXiv preprint arXiv:2508.17298, 2025.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. arXiv preprint arXiv:2408.03326, 2024a.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. arXiv preprint arXiv:2407.07895, 2024b.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 22195–22206, 2024c.

Teng Li, Quanfeng Lu, Lirui Zhao, Hao Li, Xizhou Zhu, Yu Qiao, Jun Zhang, and Wenqi Shao. Unifork: Exploring modality alignment for unified multimodal understanding and generation. arXiv preprint arXiv:2506.17202, 2025.

Xinhao Li, Yi Wang, Jiashuo Yu, Xiangyu Zeng, Yuhan Zhu, Haian Huang, Jianfei Gao, Kunchang Li, Yinan He, Chenting Wang, Yu Qiao, Yali Wang, and Limin Wang. Videochat-flash: Hierarchical compression for long-context video modeling. In The Fourteenth International Conference on Learning Representations, 2026. URL https://openreview.net/forum?id=MUjdNcfNPv.

Weixin Liang, LILI YU, Liang Luo, Srini Iyer, Ning Dong, Chunting Zhou, Gargi Ghosh, Mike Lewis, Wen tau Yih, Luke Zettlemoyer, and Xi Victoria Lin. Mixture-of-transformers: A sparse and scalable architecture for multi-modal foundation models. Transactions on Machine Learning Research, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=Nu6N69i8SB.

Bill Yuchen Lin, Ronan Le Bras, Kyle Richardson, Ashish Sabharwal, Radha Poovendran, Peter Clark, and Yejin Choi. Zebralogic: On the scaling limits of llms for logical reasoning. arXiv preprint arXiv:2502.01100, 2025.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual Instruction Tuning, December 2023. URL http://arxiv.org/abs/2304.08485. arXiv:2304.08485 [cs].

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 26296–26306, 2024.

Xiangrui Liu, Yan Shu, Zheng Liu, Ao Li, Yang Tian, and Bo Zhao. Video-xl-pro: Reconstructive token compression for extremely long video understanding, 2025. URL https://arxiv.org/abs/2503.18478.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In International Conference on Learning Representations, volume 2024, pp. 23439–23554, 2024.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 11–20, 2016.

OpenAI. gpt-5-system-card, 2025.

Roni Paiss, Ariel Ephrat, Omer Tov, Shiran Zada, Inbar Mosseri, Michal Irani, and Tali Dekel. Teaching CLIP to Count to Ten. arXiv preprint arXiv:2302.12066, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In International conference on machine learning, pp. 8748–8763. PmLR, 2021.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. Advances in Neural Information Processing Systems, 37:8612–8642, 2024.

Tao Shen, Xin Wan, Taicai Chen, Rui Zhang, Junwen Pan, Dawei Lu, Fanding Lei, Zhilin Lu, Yunfei Yang, Chen Cheng, et al. Mammothmoda2: A unified ar-diffusion framework for multimodal understanding and generation. arXiv preprint arXiv:2511.18262, 2025.

Weijia Shi, Xiaochuang Han, Chunting Zhou, Weixin Liang, Xi Victoria Lin, Luke Zettlemoyer, and LILI YU. LMFusion: Adapting pretrained language models for multimodal generation. In The Thirty-ninth Annual Conference on Neural Information Processing Systems, 2026. URL https://openreview.net/forum?id=Kc1WTxZbrP.

Fatemeh Shiri, Xiao-Yu Guo, Mona Golestan Far, Xin Yu, Reza Haf, and Yuan-Fang Li. An empirical analysis on spatial reasoning capabilities of large multimodal models. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pp. 21440–21455, 2024.

Yan Shu, Zheng Liu, Peitian Zhang, Minghao Qin, Junjie Zhou, Zhengyang Liang, Tiejun Huang, and Bo Zhao. Video-xl: Extra-long vision language model for hour-scale video understanding. In Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 26160–26169, 2025.

Zihan Su, Hongyang Wei, Kangrui Cen, Yong Wang, Guanhua Chen, Chun Yuan, and Xiangxiang Chu. Generation enhances understanding in unified multimodal models via multi-representation generation. In Forty-third International Conference on Machine Learning, 2026. URL https://openreview.net/forum?id=uYaNE2Cys2.

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. arXiv preprint arXiv:2405.09818, 2024.

Meituan LongCat Team, Bin Xiao, Chao Wang, Chengjiang Li, Chi Zhang, Chong Peng, Hang Yu, Hao Yang, Haonan Yan, Haoze Sun, et al. Longcat-next: Lexicalizing modalities as discrete tokens. arXiv preprint arXiv:2603.27538, 2026.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai C Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. Advances in Neural Information Processing Systems, 37: 87310–87356, 2024.

Shengbang Tong, David Fan, Jiachen Li, Yunyang Xiong, Xinlei Chen, Koustuv Sinha, Michael Rabbat, Yann LeCun, Saining Xie, and Zhuang Liu. Metamorph: Multimodal understanding and generation via instruction tuning. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 17001–17012, 2025.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. arXiv preprint arXiv:2502.14786, 2025.

Dianyi Wang, Wei Song, Yikun Wang, Siyuan Wang, Kaicheng Yu, Zhongyu Wei, and Jiaqi Wang. Autoregressive semantic visual reconstruction helps vlms understand better. In Findings of the Association for Computational Linguistics: ACL 2026, pp. 38101–38115, 2026a.

Haochen Wang, Anlin Zheng, Yucheng Zhao, Tiancai Wang, Zheng Ge, Xiangyu Zhang, and Zhaoxiang Zhang. Reconstructive visual instruction tuning. In International Conference on Learning Representations, volume 2025, pp. 14374–14399, 2025.

Jiacong Wang, Zijian Kang, Haochen Wang, LiangXiao, Ya Wang, Jiawen Li, Bohong Wu, Ran Jiao, Haiyong Jiang, ChaoFeng, and Jun Xiao. VGR: Visual grounded reasoning. In The Fourteenth International Conference on Learning Representations, 2026b. URL https://openreview.net/forum?id=kDhAiaGzrn.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. Advances in Neural Information Processing Systems, 37:95095–95169, 2024a.

Xinlong Wang, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Zhen Li, Yuqi Wang, Qiying Yu, Yingli Zhao, Yulong Ao, Xuebin Min, Chunlei Men, Boya Wu, Bo Zhao, Bowen Zhang, Liangdong Wang, Guang Liu, Zheqi He, Xi Yang, Jingjing Liu, Yonghua Lin, Zhongyuan Wang, and Tiejun Huang. Multimodal learning with next-token prediction for large multimodal models. Nature, 650(8101):327–333, February 2026c. ISSN 1476-4687. doi: 10.1038/s41586-025-10041-x.

Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, et al. Charxiv: Charting gaps in realistic chart understanding in multimodal llms. Advances in Neural Information Processing Systems, 37:113569–113697, 2024b.

Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. In Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 12966–12977, 2025.

Yijia Xiao, Edward Sun, Tianyu Liu, and Wei Wang. Logicvista: Multimodal llm logical reasoning benchmark in visual contexts. arXiv preprint arXiv:2407.04973, 2024.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. In The Thirteenth International Conference on Learning Representations, 2025. URL https://openreview.net/forum?id=o6Ynz6OIQ6.

Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. Show-o2: Improved native unified multimodal models. In The Thirty-ninth Annual Conference on Neural Information Processing Systems, 2026. URL https://openreview.net/forum?id=7VMg7Jb7AL.

Sihan Xu, Ziqiao Ma, Wenhao Chai, Xuweiyi Chen, Weiyang Jin, Joyce Chai, Saining Xie, and Stella X Yu. Next-embedding prediction makes strong vision learners. arXiv preprint arXiv:2512.16922, 2025.

Weiye Xu, Jiahao Wang, Weiyun Wang, Zhe Chen, Wengang Zhou, Aijun Yang, Lewei Lu, Houqiang Li, Xiaohua Wang, Xizhou Zhu, Wenhai Wang, Jifeng Dai, and Jinguo Zhu. Visulogic: A benchmark for evaluating visual reasoning in multi-modal large language models. In The Fourteenth International Conference on Learning Representations, 2026. URL https://openreview.net/forum?id=mXuzDDVXxi.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. arXiv preprint arXiv:2505.09388, 2025.

Ling Yang, Ye Tian, Bowen Li, Xinchen Zhang, Ke Shen, Yunhai Tong, and Mengdi Wang. Mmada: Multimodal large diffusion language models. Advances in Neural Information Processing Systems, 38:138867–138907, 2026.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 9556–9567, 2024.

Yichen Zhang, Da Peng, Zonghao Guo, Zijian Zhang, Xuesong Yang, Tong Sun, Shichu Sun, Yidan Zhang, Yanghao Li, Haiyan Zhao, et al. Cheers: Decoupling patch details from semantic representations enables unified multimodal comprehension and generation. arXiv preprint arXiv:2603.12793, 2026a.

YiFan Zhang, Xingyu Lu, Shukang Yin, Chaoyou Fu, Wei Chen, Xiao Hu, Bin Wen, Kaiyu Jiang, Changyi Liu, Tianke Zhang, Haonan fan, Kaibing Chen, Jiankang Chen, Haojie Ding, Kaiyu Tang, Zhang Zhang, Liang Wang, Fan Yang, Tingting Gao, and Guorui Zhou. Thyme: Think beyond images. In The Fourteenth International Conference on Learning Representations, 2026b. URL https://openreview.net/forum?id=gCWLkqK450.

Yuhui Zhang, Alyssa Unell, Xiaohan Wang, Dhruba Ghosh, Yuchang Su, Ludwig Schmidt, and Serena Yeung-Levy. Why are visually-grounded language models bad at image classification? Advances in Neural Information Processing Systems, 37:51727–51753, 2024.

Chunting Zhou, LILI YU, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. In The Thirteenth International Conference on Learning Representations, 2025. URL https://openreview.net/forum?id=SI2hI0frk6.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. arXiv preprint arXiv:2504.10479, 2025.

Chengke Zou, Xingang Guo, Rui Yang, Junyu Zhang, Bin Hu, and Huan Zhang. Dynamath: A dynamic visual benchmark for evaluating mathematical reasoning robustness of vision language models. In International Conference on Learning Representations, volume 2025, pp. 48337–48383, 2025.

A Related Works and Comparison with Previous Works

Multimodal large language models (MLLMs) have emerged as the prevailing paradigm for image and video understanding, typically by aligning pretrained visual encoders with powerful language backbones. Representative early systems include Flamingo Alayrac et al. (2022) and InstructBLIP Dai et al. (2023), whereas subsequent open-source families such as the LLaVA series Li et al. (2024a); Liu et al. (2023); Li et al. (2024b) and Qwen-VL series Bai et al. (2025a;b) have further advanced image perception and understanding. Most studies in this line largely follow the LLaVA paradigm Liu et al. (2023), in which visual inputs are first encoded by a vision encoder Radford et al. (2021); Tschannen et al. (2025) and subsequently concatenated with text tokens for joint modeling by a language model decoder. Proprietary systems such as GPT OpenAI (2025) and Gemini Google (2024; 2025) likewise exhibit strong multimodal understanding capabilities. Recent advances have further extended these models to multimodal visual reasoning Xu et al. (2026); Wang et al. (2026b); Ke et al. (2025) and video understanding Liu et al. (2025); Shu et al. (2025); Li et al. (2026). Despite their notable strengths in semantic abstraction and cross-modal alignment, these models are primarily optimized for text generation rather than learning from native visual information. The training signal flows exclusively from vision to language – visual tokens serve as a conditioning context for text prediction, but no supervisory gradient encourages the model to develop fine-grained, pixel-level visual perception or to retain visual evidence throughout deep reasoning chains. This asymmetry limits performance on tasks that demand precise spatial understanding, such as grounding, segmentation, and complex visual chain-of-thought reasoning.

Table 8: **Design comparison of representative approaches that combine visual understanding and generation.** The attributes characterize how generation is used during training rather than absolute model quality. GAS is the only compared design that combines all six properties.

Method	Generation target	U-centric	Task-cond.	Exact U-space	No ext. codec	Upper U-isol.	Train-only aux.
Native unified understanding–generation models							
Chameleon Team (2024)	Discrete image tokens	✗	✗	✗	✗	✗	✗
Show-o Xie et al. (2025)	Discrete diffusion tokens	✗	✓	✗	✗	✗	✗
Janus-Pro Chen et al. (2025c)	Discrete image tokens	✗	✗	✗	✗	✗	✗
MetaMorph Tong et al. (2025)	Vision-encoder features	●	✓	✗	✗	✗	✗
BAGEL Deng et al. (2025)	VAE/semantic features	✗	✓	✗	✗	●	✗
UniFork Li et al. (2025)	Discrete image tokens	✗	✗	✗	✗	✓	✗
Generation supervision designed to enhance understanding							
ROSS Wang et al. (2025)	Continuous VAE features	✓	✗	✗	✗	✗	✓
ASVR Wang et al. (2026a)	Discrete semantic tokens	✓	✗	✗	✗	✗	✓
UniHetero Chen et al. (2025a)	LLM-input embeddings	✓	✗	✓	✗	✗	●
UniMRG Su et al. (2026)	RGB/depth/segmentation	✓	✓	✗	✗	●	✗
GAS (Ours)	**LLM-input embeddings**	✓	✓	✓	✓	✓	✓

Definitions. **U-centric**: generation is explicitly optimized to improve understanding rather than retained as a co-equal capability; **Task-cond.**: an observed visual input and instruction determine a distinct target, beyond generic text-to-image synthesis or input reconstruction; **Exact U-space**: the target is the projected continuous embedding consumed by the understanding LLM; **No ext. codec**: no separate visual tokenizer, VAE, or diffusion decoder defines the generation target; **Upper-U isolated**: generation gradients do not update the upper understanding transformer; **Train-only aux.**: generation-specific machinery is introduced for understanding training and removed at deployment. ● denotes partial or architecture-dependent support.

Recent unified multimodal models (UMMs) attempt to close this gap by jointly training understanding and generation within a single framework, thereby introducing visual synthesis objectives that force

the model to engage more deeply with image structure. One line follows a fully autoregressive formulation: Chameleon Team (2024) and Emu3 Wang et al. (2026c) cast both understanding and generation into next-token prediction over a shared discrete visual token space, offering a clean interface for mixed-modality sequence modeling. Another line adopts AR-diffusion hybrid formulations, combining language modeling for text with diffusion or flow-based objectives for visual generation. Representative works include Transfusion Zhou et al. (2025), Show-o Xie et al. (2025; 2026), and BLIP3-o Chen et al. (2025b). Within this family, several works further explore architectural decoupling to mitigate the well-known conflict between generation and understanding objectives: Janus Chen et al. (2025c) decouples visual encoding pathways while sharing a single LLM backbone; MetaMorph Tong et al. (2025) introduces Visual-Predictive Instruction Tuning and observes that improving visual understanding implicitly benefits generation; LMFusion Shi et al. (2026) adds modality-specific transformer layers to preserve pretrained language capabilities during generation training; and BAGEL Deng et al. (2025) scales on trillions of interleaved tokens and exhibits compositional generation and reasoning.

Most closely related to our goal, ROSS Wang et al. (2025) reconstructs the input image through a denoising objective over continuous appearance features, whereas ASVR Wang et al. (2026a) autoregressively reconstructs discrete semantic visual tokens. UniHetero Chen et al. (2025a) further shows at large data scale that semantic generation—including autoregression on the LLM input embedding—can improve understanding, while pixel-level generation gradients can be detrimental. UniMRG Su et al. (2026) post-trains existing UMMs to generate RGB reconstructions, depth maps, and segmentation maps from an input image. These results strengthen the broad premise that visual-side supervision can improve understanding; accordingly, GAS's distinction is the combination of instruction-conditioned output-image prediction across diverse generation tasks, direct supervision in the understanding input space, asymmetric upper-layer isolation, and a generation branch used only during training.

Appendix Table 10 further compares adapted ROSS-style and ASVR-style objectives under the same backbone, data mixture, and optimization setting. These controls test how their reconstruction objectives transfer to our generation-augmented setting rather than replacing the original papers' reported results.

However, despite these architectural innovations, existing UMMs predominantly pursue generation quality or explore emergent capabilities from scale, while treating the understanding branch as a secondary beneficiary rather than the primary optimization target. Moreover, the introduction of generation objectives frequently degrades understanding performance due to conflicting gradient signals, a tension that current methods address by either accepting the tradeoff or increasing model capacity Team et al. (2026). Currently, there exists no systematic investigation into methods for integrating visual generation supervision to preserve and further advance models' semantic understanding capabilities.

B Detailed Generation Tasks

This section provides detailed descriptions of the five generation task categories and their 15 subtasks, including Grounding, Segmentation, Image Editing, Visual Chain-of-Thought, and Text-to-Image (T2I).

- **Grounding**: Region localization tasks that compel the model to perfectly resolve complex linguistic referring expressions before predicting absolute spatial coordinates, directly cultivating fine-grained spatial reasoning.
- **Segmentation**: Pixel-level mask prediction tasks that force the model to comprehend semantic boundaries and physical extents based on intricate textual instructions, reinforcing instance-level discrimination.
- **Image Editing**: Conditional image transformation tasks (e.g., style transfer, object removal) that require a deep, compositional understanding of scene layouts and object interdependencies before executing the edit.

- **Visual Chain-of-Thought**: Multi-step reasoning tasks that demand the model to actively generate intermediate visual evidence (e.g., highlighting key regions) to anchor multi-step logical deduction, simulating a "thinking with images" process.
- **Text-to-Image (T2I)**: Carefully curated image synthesis tasks grounded in rich world knowledge and dense structural prompts, establishing foundational cross-modal alignment beyond naive concept mapping.

C Detailed Evaluation Benchmarks

C.1 General Perception

We evaluate holistic multimodal understanding and multi-image reasoning on MME Fu et al. (2026a), MMMU Yue et al. (2024), BLINK Fu et al. (2024), and RealWorldQA. These benchmarks assess broad visual perception, world knowledge integration, and multi-image comprehension.

C.2 Visual Reasoning

We evaluate mathematical and logical visual reasoning on CharXiv Wang et al. (2024b), Dyna-Math Zou et al. (2025), MathVision Wang et al. (2024a), MathVista Lu et al. (2024), LogicVista Xiao et al. (2024), and VisuLogic Xu et al. (2026). These benchmarks require precise visual parsing of charts, diagrams, and geometric figures, together with multi-step logical inference.

C.3 Counting and Spatial Understanding

We evaluate fine-grained spatial perception using CountBench Paiss et al. (2023); Beyer et al. (2024) and CV-Bench Tong et al. (2024). These benchmarks specifically probe object counting accuracy and spatial relationship understanding, which are closely related to the capabilities trained by our grounding and segmentation tasks.

C.4 Video Understanding

We evaluate temporal reasoning on Video-MME Fu et al. (2025) and MVBench Li et al. (2024c). These benchmarks assess sustained visual attention across frames and the ability to perform temporal reasoning, thereby testing whether the benefits of generation-guided training generalize to the video domain.

D Additional Controlled Experiments

D.1 Repeated-Run Robustness Across Data Scales

We repeat GAS three times in the 2.5B-token regime and twice in the 20B-token regime. Table 9 reports the sample mean and sample standard deviation, together with the original and longer-trained understanding-only controls. The GAS mean remains above both single-run controls in Overall at each scale. Variation is larger for individual capabilities—most notably Reasoning at 2.5B and Count&Spatial at 20B—so these runs support reproducible aggregate transfer.

D.2 Aligned Comparison with Reconstructive Objectives

We reproduce ROSS-style and ASVR-style objectives using the same Qwen3-VL-2B backbone, 2.5B understanding plus 2.5B generation data, and optimization setting as GAS. On generation samples, these adapted controls reconstruct the assistant output image using continuous appearance features and discrete semantic tokens, respectively. As shown in Table 10, input-reconstruction objectives designed for understanding data do not automatically transfer to this generation-augmented setting, whereas GAS's task-conditioned output prediction gives the strongest aggregate result. These are controlled adaptations for isolating the objective under our setup, rather than reproductions of the original papers' reported training regimes.

Table 9: **Repeated-run robustness across training scales.** GAS entries are mean ± sample standard deviation; baseline controls are single runs. Bold marks the best reported value within each regime and capability.

Regime	Configuration	Runs	Overall	Perception	Reasoning	Count&Spatial	Video
2.5B	Baseline	1	47.25	56.55	31.38	**76.92**	46.60
2.5B	Longer baseline (2.77B; +11%)	1	47.73	57.63	**34.01**	71.28	45.55
2.5B	GAS	3	48.46±0.37	58.56±0.46	32.99±1.34	75.83±1.06	47.28±0.35
20B	Baseline	1	51.95	60.12	38.88	79.62	47.10
20B	Longer baseline (22B; +10%)	1	52.14	60.18	38.58	**80.90**	46.95
20B	GAS	2	52.33±0.33	61.04±0.45	39.05±0.24	80.71±1.54	47.35±0.49

Table 10: **Same-backbone comparison with adapted reconstructive objectives.** All U+G methods use the same backbone, data mixture, and optimization setting.

Method	Overall	Perception	Reasoning	Count&Spatial	Video
Understanding-only baseline	47.25	56.55	31.38	**76.92**	46.60
ROSS-style (U+G)	45.70	56.10	30.88	70.80	44.25
ASVR-style (U+G)	47.31	56.20	**33.16**	73.78	45.55
GAS (U+G)	**48.25**	**58.52**	32.63	75.15	**47.65**

E More Attention Map Analysis Comparison

On MathVista object-counting tasks, GAS correctly attends to all relevant objects for subtraction reasoning, whereas the baseline's scattered attention leads to miscounting.

Qualitative Case 2: MathVista-MINI (Attention Analysis)

Question: Subtract all red things. Subtract all tiny matte balls. How many objects are left?
Ground Truth: 5
Baseline: **<think>** Identifies the red matte ball as both "red" and "tiny matte ball", but then double-counts the remaining items, mistakenly listing 6 leftover objects. **</think>** Answer: **6** ✗
GAS: **<think>** Total 7 objects; the red sphere is the only object that is both red and a tiny matte ball, so it is removed once. Remaining objects: yellow cube, green small cylinder, cyan sphere, blue / green / purple large cylinders, giving $7 - 1 - 1 = 5$. **</think>** Answer: **5** ✓

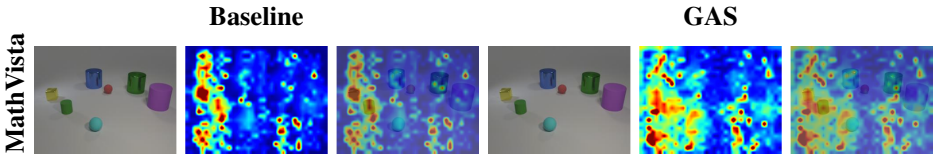
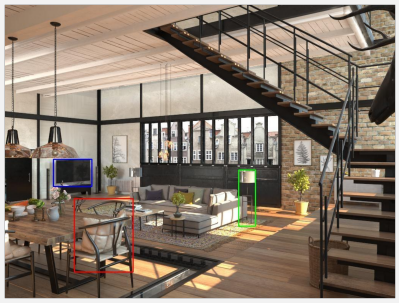


Figure 6: **Attention visualization on representative understanding samples in MathVista Lu et al. (2024).** The GAS model produces the corrected predictions shown in Boxes 2.

F MORE QUALITATIVE CASES OF GAS

Qualitative Case 3: CV-Bench Tong et al. (2024)



Question: Estimate the real-world distances between objects in this image. Which object is closer to the chair (highlighted by a red box), the television (highlighted by a blue box) or the lamp (highlighted by a green box)? A. television, B. lamp

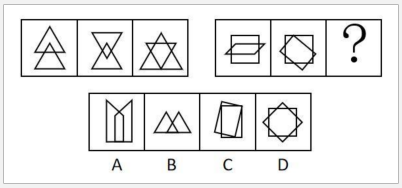
---

GAS-4B:

<think>
To determine which object is closer to the chair (highlighted by the red box), we need to analyze the spatial relationships in the image:
1. **Chair (red box)**: This is located in the foreground, near the dining table.
2. **Television (blue box)**: This is positioned on a stand against the wall, further back in the room.
3. **Lamp (green box)**: This is placed on the floor near the sofa, closer to the center of the room but still farther back than the chair.
Analysis: - The **chair** is in the foreground, very close to the viewer.
- The **television** is on a stand against the wall, which is significantly farther away from the chair.
- The **lamp** is on the floor near the sofa, which is also farther away from the chair compared to the chair's position.
Conclusion: - The **chair** is the closest object to itself, and the **lamp** is closer to the chair than the **television**.
Thus, the correct answer is:
B. lamp
</think>

The final answer is $\boxed{B}$.

Qualitative Case 4: VisuLogic Xu et al. (2026)



Question: Choose the most appropriate one from the four given options to fill in the question mark so that a certain regularity is presented: A, B, C, D.

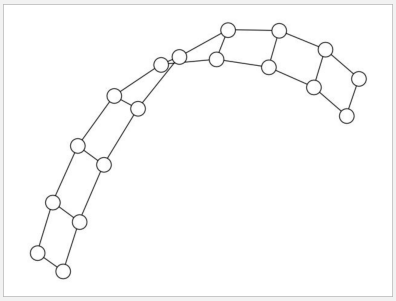
Ours 4B:

<think>
To solve this problem, we first analyze the shapes and number of lines in each figure, then identify the underlying pattern.
1. **Analyze shape and line count:**
Observing the figures in the first and second rows, each figure is composed of **triangles** or **quadrilaterals** (rhombus-like shapes), with a regular line count.
- Row 1, Fig. 1: two triangles, lines $= 2 \times 3 = 6$;
- Row 1, Fig. 2: two quadrilaterals, lines $= 2 \times 4 = 8$;
- Row 1, Fig. 3: two triangles, lines $= 2 \times 3 = 6$;
- Row 2, Fig. 1: two quadrilaterals, lines $= 2 \times 4 = 8$;
- Row 2, Fig. 2: two triangles, lines $= 2 \times 3 = 6$;
- Row 2, Fig. 3 (the question mark): should follow the same shape/line-count pattern and stacking style.
2. **Identify the pattern:**

The figures in both rows alternate between **triangle pairs** and **quadrilateral pairs**, with line counts of 6 and 8 respectively. All figures share the same two-shape overlapping composition with consistent contour and inner-line connectivity.
3. **Match the options:**
Option D consists of two overlapping quadrilaterals (rhombus-like), whose line count and composition match the pattern (Row 2 starts with a quadrilateral pair, so the question mark should also be a quadrilateral pair). Option A is a triangle pair with a mismatched shape; Option B is a triangle pair with inconsistent line count; Option C is a quadrilateral but with a different stacking style.
Therefore, the correct answer is $\boxed{D}$.
</think>

The correct answer is $\boxed{D}$.

Qualitative Case 5: DynaMath Zou et al. (2025)

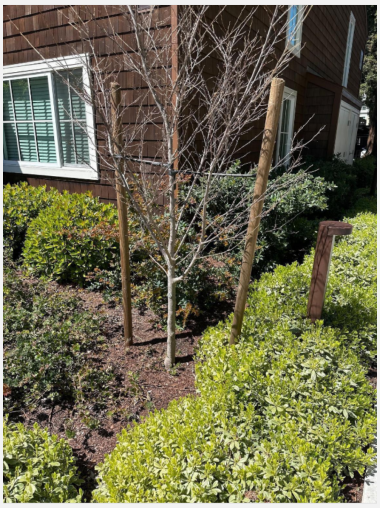


Question: Find the chromatic number of the following graph.

Ours 4B:
<think> To determine the chromatic number of the given graph, we need to find the minimum number of colors required to color the vertices such that no two adjacent vertices share the same color. Let's analyze the graph step by step.
1. **Identify the Structure of the Graph:** - The graph appears to be a bipartite graph. A bipartite graph is a graph whose vertices can be divided into two disjoint sets such that every edge connects a vertex in one set to a vertex in the other set. - In this graph, we can observe that the vertices can be divided into two sets: one set containing the vertices on the left side of the graph and the other set containing the vertices on the right side of the graph.
2. **Verify Bipartiteness:** - Check if there are any odd-length cycles in the graph. If there are no odd-length cycles, the graph is bipartite. - Upon inspection, the graph does not contain any odd-length cycles. All cycles in the graph are even-length cycles.
3. **Determine the Chromatic Number:** - For a bipartite graph, the chromatic number is always 2. This is because we can color one set of vertices with one color and the other set with a different color, ensuring that no two adjacent vertices share the same color.
4. **Conclusion:** - Since the graph is bipartite, the chromatic number is 2.
</think>

The final answer is $\boxed{2}$.

Qualitative Case 6: RealWorldQA



Question: How many wooden poles are in this photo?

Ours 4B:

<think>
To determine the number of wooden poles in the photo, let's analyze the image step by step:
1. **Foreground Analysis**: The image shows a small tree with a thin trunk. There are two visible wooden poles supporting the tree. These poles are positioned on either side of the tree, providing stability.
2. **Background and Surroundings**: The background includes a building with windows and some greenery. There are no additional wooden poles visible in the background or elsewhere in the image.
3. **Counting the Poles**: Based on the visible elements, there are exactly two wooden poles supporting the tree.
Given this analysis, the correct answer is:
C. There are 2 wooden poles.
</think>

The final answer is $\boxed{C}$.